Novas perspetivas do *Business Intelligence*: criação de novos indicadores de desempenho

por

Tânia Patrícia Serra Veloso

Dissertação de Mestrado em Economia e Administração de Empresas

Orientada por

Professor Doutor João Pedro Carvalho Leal Mendes Moreira

Professor Doutor Pedro José Ramos Moreira de Campos

- 2017 –

NOTA BIOGRÁFICA

A 1 de Outubro de 1993, nascia, em Vila Nova de Famalicão, Tânia Patrícia Serra Veloso.

No ano letivo de 2011/12, inicia a sua experiência académica ao ingressar na Licenciatura de Economia da Universidade do Minho. Daqui, sairia com destino à Universidade de Pardubice, República Checa, onde realizou Erasmus no último ano da Licenciatura.

No Verão de 2014, concluída a Licenciatura, inaugura uma nova etapa na cidade do Porto, onde ingressa no Mestrado em Economia e Administração de Empresas da Faculdade de Economia do Porto.

Praticamente em simultâneo, inicia a sua carreira profissional na Mota-Engil Serviços Partilhados. Empresa onde, após realização de estágio e passagens por diversos departamentos, ainda hoje se encontra como responsável pelas Contas a Pagar para os mercados da América Latina.

AGRADECIMENTOS

Depois de alcançada mais uma etapa tão importante como esta, cabe-me agradecer de forma especial a todos os que me acompanharam e acreditaram em mim ao longo da elaboração desta dissertação.

Ao meu orientador, Professor Doutor João Moreira e ao meu coorientador Professor Doutor Pedro Campos por todo o apoio que foi para além do tremendo conhecimento técnico. Agraço assim, também por toda a disponibilidade, paciência, confiança e motivação demonstrada ao longo deste percurso. Foram, evidentemente, um pilar fundamental para a concretização deste projeto.

Um agradecimento sentido ao Vasco Veludo e à Magda Rocha pelo carinho e motivação que me transmitiram sempre ao longo deste longo processo.

E claro, aos meus pais e irmão por todo o apoio incondicional, não só na fase de elaboração desta dissertação, como desde sempre.

RESUMO

Esta dissertação trata um problema atual de *Business Intelligence* (BI) em que se procura definir novos indicadores de desempenho para efeitos de tomada de decisão. Habitualmente, os indicadores de desempenho são somáveis ou, pelo menos, agregáveis de forma relativamente fácil (por exemplo, o número de clientes que fazem compras nas lojas, ou os lucros de uma empresa). Com a elaboração deste estudo, pretende-se introduzir uma inovação, através da aplicação simultânea de técnicas de C*lustering* e Regressão para definir indicadores de performance em situações em que certos indicadores não são somáveis (por exemplo, o índice de poder de compra, ou a taxa de desemprego).

Para o efeito, foi usada uma base de dados contendo informação sócio económica dos municípios e regiões de Portugal. Através dessa base, foram construídas regressões lineares que, por sua vez, foram sujeitas a um processo de agrupamento, ou *Clustering.* O agrupamento mencionado foi realizado de modo a que existam determinadas semelhanças entre as regressões pertencentes a cada um dos grupos com o objetivo de facilitar a interpretação e análise dos dados. Deste modo, por exemplo, podem-se agregar os municípios em regiões ou desagregar regiões em níveis com maior granularidade, consoante pretenda obter uma visão mais alargada e abrangente dos dados ou mais específica e pormenorizada, em situações onde a simples soma (agregação) ou subtração (desagregação) de indicadores não faz sentido.

Palavras-chave: *Business Intelligence*, Indicadores de Desempenho, Tomada de Decisão, C*lustering*, Regressões Lineares, Agregação e Desagregação.

ABSTRACT

This dissertation addresses a current Business Intelligence (BI) problem in which new performance indicators are sought for decision-making purposes. Usually, performance indicators are summable or at least relatively easily aggregable (for example, the number of shoppers, or the profits of a company). With the elaboration of this study, we intend to introduce an innovation, through the simultaneous application of Clustering and Regression techniques to define performance indicators in situations where certain indicators are not summable (for example, the purchasing power index, or unemployment rate).

For this purpose, a database containing socioeconomic information was used for the municipalities and regions of Portugal. Through this base, linear regressions were constructed which, in turn, were subjected to a grouping process, or Clustering. The grouping was performed in such a way that there are certain similarities between the regressions belonging to each of the groups in order to facilitate the interpretation and analysis of the data. In this way, for example, municipalities can be aggregated into regions or disaggregated into more granular levels, depending on whether they want a broader and more comprehensive view of the data or more specific and detailed, in situations where simple sum (aggregation) or subtraction (disaggregation) of indicators does not make sense.

Keywords: Business Intelligence, Performance Indicators, Decision-making, Clustering, Linear Regressions, Aggregation e Disaggregation.

ÍNDICE DE CONTEÚDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

ABREVIATURAS

BI - *Business Intelligence*

INE - Instituto Nacional de Estatística

KPI - *Key Performance Indicator*

NUTS - Nomenclatura das Unidades Territoriais para Fins Estatísticos

OLAP - *Online Analytical Processing*

OLTP - *Online Transaction Processing*

UE- União Europeia

VAB - Valor Acrescentado Bruto

1. INTRODUÇÃO

1.1. Enquadramento e Motivação

Numa era de forte desenvolvimento tecnológico, associada a uma concorrência cada vez mais intensa entre organizações, a área da Inteligência de Negócio, ou *Business Intelligence* (BI), tem assumido particular preponderância no mundo empresarial. Esta área dedica-se ao estudo de dados ao dispor de uma organização, tendo como objetivo fornecer informação relevante, com vista a tornar os procedimentos mais eficazes.

Nos últimos tempos, temos visto um crescimento explosivo, tanto do número de produtos e serviços oferecidos pelas empresas, como da adoção das tecnologias por parte do mundo empresarial. Em particular, têm-se desenvolvido armazéns de dados (*Data Warehouses*) a partir dos quais assentam processos de inteligência de negócio. O armazenamento e gestão de dados nas empresas são fundamentais na tomada de decisões, permitindo que o utilizador (executivo, gerente, analista) possa tomar melhores e mais rápidas decisões. Chaudhuri e Dayal (1997).

Com efeito, a área de *Business Intelligence* está regularmente associada a armazéns de dados. Estes contemplam a informação interna e externa que possa ser objeto de análise para, posteriormente, auxiliar a tomada de decisão. Porém, existem diversas formas para analisar os dados referidos. Deste modo, também associadas ao *Business Intelligence*, estão as tecnologias utilizadas para retirar os dados dos *Data Warehouses* e tratá-los de modo a serem de fácil leitura para os responsáveis pela tomada de decisão.

Uma das tecnologias comumente utilizada para efetuar a análise é o *Data Mining*. Através desta, torna-se possível identificar padrões e criar relações entre a enorme quantidade de dados alojada nos supracitados *Data Warehouses*. A tecnologia de *Data Mining* é vasta e constituída por diversas técnicas.

As técnicas de extração de dados (a partir do D*ata Warehouse*) e análise de dados (*Data Mining*) dependem do tipo de informação que é relevante para o negócio. A criação de indicadores de gestão a partir desses dados é fundamental, mas alguns desses indicadores não se encontram otimizados para fornecer visões interessantes sobre os dados.

Na prática, o que é proposto neste trabalho é a aplicação de uma técnica de *Data Mining* específica, as regressões lineares múltiplas diversas como forma de fornecer indicadores de gestão. Posteriormente, utiliza-se uma outra técnica de *Data Mining*, o C*lustering.* Este, será aplicado sobre indicadores calculados através das regressões lineares múltiplas.

Mais especificamente, apesar da técnica de C*lustering* estar já bastante desenvolvida, esta aplica-se, usualmente, ao agrupamento de dados individuais. Porém, o intuito desta dissertação será a aplicação de C*lustering* não a dados individuais, mas sim a regressões lineares múltiplas com o intuito de as agrupar num número de regressões que seja humanamente possível de analisar.

1.2. Abordagem

A sequência de tarefas levadas a cabo para a realização deste trabalho iniciou-se, como é habitual, pela escolha do tema. Após a definição da área sobre a qual a dissertação se iria debruçar, iniciou-se uma pesquisa bibliográfica com recurso a artigos científicos, teses já realizadas, livros e diversas publicações com a finalidade de definir uma área concreta dentro da área de *Business Intelligence*.

A escolha de uma área de estudo revelou-se complexa, tendo em conta a enorme diversidade de temas associados ao *Business Intelligence*. Porém, através da pesquisa realizada, optou-se por abordar um tema ainda pouco explorado. Apesar de tornar a pesquisa mais aliciante e, por seu turno, a realização do trabalho, esse mesmo fator aumenta o desafio.

Tendo em conta o facto de o foco da dissertação juntar duas áreas distintas, seria pertinente abordar o assunto em separado, de modo a tornar percetível as bases sobre as quais a construção do trabalho iria evoluir. Assim sendo, a pesquisa bibliográfica posterior incidiu sobre dois conjuntos distintos. Por um lado, o *Business Intelligence*, a sua definição e os seus conteúdos, bem como as áreas que aborda. No fundo, uma continuação, embora mais vocacionada, da pesquisa previamente efetuada para a definição do tema. Com efeito, neste caso, a pesquisa já foi mais direcionada para as diferentes formas de agrupamento de dados. A outra parte da pesquisa bibliográfica foi, ao mesmo tempo, uma revisão de matéria anteriormente aprendida. Refiro-me às regressões lineares, tema central nesta dissertação e, por isso mesmo, especialmente importante de ficar bem definido e explicado.

A junção dos dois temas descritos é, na realidade, o ponto-chave do estudo realizado. Após esta fase teórica, o estudo necessitava de ser testado empiricamente. Para tal, foi necessário definir uma base de dados que permitisse a aplicação do método pretendido. Essa base tinha de ser passível de agregar e desagregar, bem como de conter indicadores complexos mas não somáveis, permitindo o cálculo das regressões lineares e consequente C*lustering*.

1.3. Estruturação da Dissertação

Na abordagem aplicada nesta revisão de literatura, optou-se por familiarizar o leitor com diversos conceitos e termos que, devido à sua complexidade técnica, possam não ser de fácil compreensão inicial. Porém, esta aproximação geral é essencial para a compreensão dos passos seguintes onde a especificação será mais elevada. Assim sendo, no Capítulo 1, o trabalho começa pela introdução ao *Business Intelligence*, e *Data Mining*.

No Segundo Capítulo, serão abordadas especificamente as técnicas de regressão linear múltiplas e o *Clustering*. Seguidamente, será descrita a forma como se pretendem interligar ambas as técnicas com particular destaque para a importância da existência de indicadores complexos. Esta perspetiva acerca do estudo de indicadores e aplicabilidade dos mesmos é algo original e pouco explorado. Este facto, traz um maior interesse ao estudo do tema e torna-o bastante mais desafiante, nunca esquecendo a responsabilidade que também acarreta o tratamento de algo pouco estudado a nível científico.

De seguida, no Capítulo 3, apresenta-se o estudo do caso. Recorrendo a situações reais, aborda-se a diferença entre a utilização de dados somáveis e de regressões. Optou-se por iniciar o estudo de caso com a explicação para o uso de Regressões e não de dados somáveis, como uma breve introdução que esclareça o leitor acerca da vantagem na utilização deste modelo. Seguidamente, passa-se para a análise de dados, na qual se começa por utilizar um caso com dados somáveis e como estes são tratados. Nesta parte do trabalho, explicam-se os indicadores não somáveis que são utilizados, bem como a forma como estes são tratados, calculando as regressões lineares múltiplas. Tal como já foi explicado, o passo seguinte é a concretização do C*lustering* que irá permitir o agrupamento das regressões. Por último, aplicam-se os coeficientes de determinação que dão a conhecer a qualidade do modelo criado.

No Quarto e último Capítulos da Dissertação, retiram-se as conclusões finais e apresenta-se o contributo que este estudo traz para o *Business Intelligence.* Por sua vez, também se apresentam as limitações do estudo efetuado e sugestões de investigação futura.

Por fim, segue-se a Bibliografia, bem como os Anexos.

2. REVISÃO DE LITERATURA

2.1. *Business Intelligence*

O *Business Intelligence* é uma componente de apoio operacional do negócio que procura recolher e gerir dados internos e externos às empresas, de modo a gerar valor acrescentado para a tomada de decisão. Numa organização a informação é um importante pilar que é usado com duas grandes finalidades: a manutenção de registos operacionais e a tomada de decisões analíticas. Na primeira finalidade, apenas se colocam os dados, registando-os. Por seu turno, através dos sistemas de *Data Warehouse* e *Business Intelligence* podem-se obter esses dados, de forma a suportar a tomada de decisão. Na realidade, com estes sistemas pode-se recolher informação diretamente sem haver necessidade de intermediários, contrariando as características do modelo racional normalizado (Kimball, 2013).

A facilidade e extrema simplificação que o *Data Warehouse* e o *Business Intelligence* oferecem fazem com que o filtro existente não seja tão eficaz, guardando informação redundante. Noutro sentido, as pesquisas efetuadas na base de dados servem para traçar um perfil e, desta forma, as pesquisas seguintes irão ser condicionadas pelas anteriores. Outra característica destes sistemas é estarem otimizados para pesquisas de base de dados. Por fim, estima-se que o espaço ocupado pela base de dados do sistema *Data Warehouse* é muito maior do que o que é ocupado pelo modelo racional normalizado (Kimball, 2013).

Segundo Chaudhuri e Dayal (1997), o armazenamento de dados é um importante suporte tecnológico à tomada de decisões, permitindo que o utilizador (executivo, gerente, analista) possa tomar melhores e mais rápidas decisões. Nos últimos três anos temos visto um crescimento explosivo, tanto do número de produtos e serviços oferecidos, como uma adoção destas tecnologias por parte da indústria.

Normalmente, o armazém de dados funciona autonomamente em relação à organização das bases de dados operacionais. O armazém de dados contém o processamento analítico *online* (OLAP). Porém, os requisitos de funcionamento e de desempenho destes, são distintos do processamento de aplicações de transações *online* (OLTP) relacionados com as bases de dados operacionais. Os sistemas OLTP têm como tarefa automatizar e

processar as funções básicas e rotineiras de uma organização. Em contrapartida, os armazéns de dados são direcionados para o apoio à decisão. Na grande maioria dos casos, os dados resumidos e consolidados, são mais fundamentais que os registos individuais e bastante pormenorizados. Para facilitar a visualização e análises complexas, os dados em armazém são normalmente apresentados em modelos multidimensionais (Chaudhuri & Dayal, 1997).

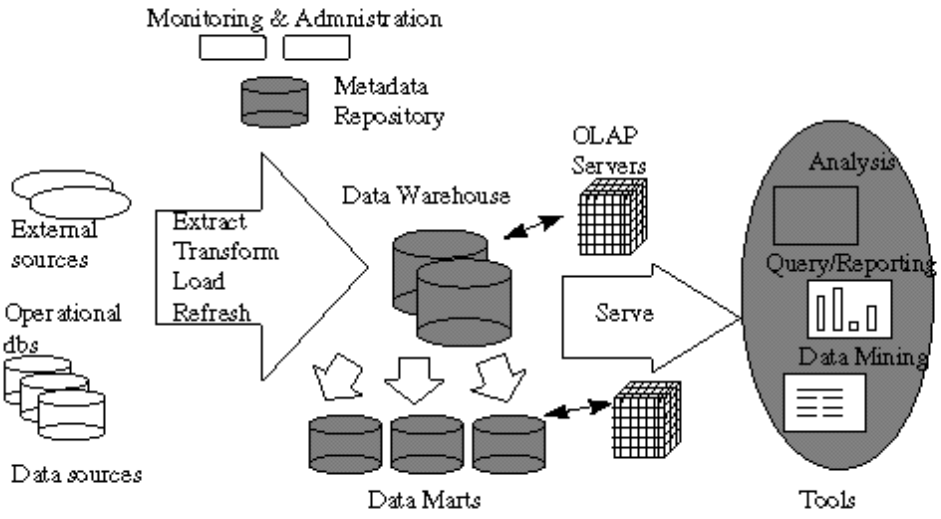


Figura 1 - Processo de *Data Warehousing*
Fonte: Chaudhuri & Dayal, 1997, pp. 2

O armazenamento de dados inclui diversas ferramentas com funções distintas. Entre elas, estão as que extraem dados de várias bases de dados operacionais e de fontes externas. Por outro lado, existem também ferramentas para limpeza, transformação e integração de dados, bem como para carregar dados para o armazém de dados. Outros tipos de ferramentas têm como finalidade atualizar periodicamente o armazém para atualizar alterações nas fontes (Chaudhuri & Dayal, 1997).

Além do depósito principal podem existir vários armazéns de dados departamentais. Os dados armazenados em repositórios globais ou em repositórios departamentais são

geridos por um ou mais servidores de armazém, que apresentam visões multidimensionais de dados para uma variedade de ferramentas de *front-end*: ferramentas de consulta, elaboração de relatórios, ferramentas de análise e ferramentas de mineração de dados. Há também um repositório para armazenar e gerir metadados e ferramentas para monitorizar e administrar o sistema de armazenamento. O armazém pode ser distribuído para balanceamento de carga e alta disponibilidade (Chaudhuri & Dayal, 1997).

2.2. *Data Mining*

Numa sociedade extremamente informatizada, o *Data Mining* surge com um papel cada vez mais preponderante na organização de dados. Na realidade, a principal função associada a esta ferramenta é a organização de quantidades extremamente elevadas de dados que, sem a utilização do *Data Mining*, seriam de complicado entendimento. De forma geral, o *Data Mining* atua como um descodificador transformando informação impercetível e, aparentemente, com pouco interesse em conteúdos com informação extremamente útil para o utilizador (Vaisman & Zimányi, 2014).

Apesar da extrema relevância do *Data Mining*, este é apenas uma das ferramentas inseridas num processo complexo de descodificação de dados. Na realidade, o que torna o *Data Mining* bastante eficaz é a sua versatilidade, tanto a nível de áreas científicas em que se baseia, como na procura de informação relevante. A inteligência artificial, tal como a estatística ou as redes neurais, são algumas das áreas em que o *Data Mining* se baseia. Deste modo, aumentam-se as fontes de informação e analisa-se um problema através de diferentes prismas, havendo uma visão mais completa do tema (Vaisman & Zimányi, 2014).

Tal como já foi referido, o processo de organização de dados e recolha de informação levado a cabo pelo *Data Mining* é extremamente complexo e engloba inúmeros dados distintos. Porém, outra vantagem associada a esta ferramenta é a forma prática como a expõe ao utilizador. Não raras vezes, o investigador que utiliza o *Data Mining* não tem perceção da ligação entre a informação que possui e os resultados aos quais quer chegar. Assim sendo, o *Data Mining* também torna a informação mais simples e intuitiva ao utilizador (Vaisman & Zimányi, 2014).

Convém salientar que este método não irá inventar nada inexistente, simplesmente torna mais percetível a informação dispersa e de difícil análise para o utilizador. O *Data Mining* funciona através de inputs que o utilizador insere e, através destes, irão ser criados outputs. Estes últimos, terão a forma de exemplos e previsões feitas com base nos dados inseridos pelos utilizadores. Com efeito, será o utilizador a especificar as categorias inseridas de dados e que valores podem tomar. Posteriormente, o *Data Mining* analisa a informação introduzida através de algoritmos (Witten & Frank, 2000).

2.2.1. O Processo de *Data Mining*

O modelo de processo de *Data Mining* CRISP-DM fornece uma visão geral do ciclo de vida de um projeto de *Data Mining.* Com efeito, o *Data Mining* não termina quando a solução é implementada. Na realidade, as lições tiradas durante o processo e a solução implementada podem levar ao surgimento de novas questões de negócios. O processo CRISP-DM consiste num ciclo de seis fases, no entanto, a sua sequência não é rígida (Chapman et al, 2000).

Na primeira fase, pretende-se conhecer melhor o negócio. Assim sendo, é importante, numa fase inicial, clarificar quais os objetivos e requisitos do projeto. Seguidamente, a partir de uma perspetiva de negócio, converter esse conhecimento num problema de *Data Mining*, juntamente com um plano preliminar projetado para atingir esses objetivos (Chapman et al, 2000).

Por seu turno, na segunda fase será necessário compreender os dados. Isto consiste em identificar as fontes com a finalidade de familiarizar-se com os dados recolhidos. Desta forma, será possível identificar problemas de qualidade nos dados e detetar subconjuntos, com a finalidade de formar hipóteses de informação escondida (Chapman et al, 2000).

A fase seguinte é a preparação dos dados. Esta, abrange todas as operações de limpeza e transformação de dados de modo a construir uma base de dados final com os dados iniciais brutos (Chapman et al, 2000).

Na quarta fase, a modelação, irão ser selecionadas várias técnicas e aplicadas. Assim, irá extrair-se o conhecimento necessário para resolver o problema de negócio (Chapman et al, 2000).

A quinta fase é a avaliação, nesta, é avaliada a adequação dos conhecimentos obtidos e dos passos executados na construção do modelo. Por sua vez, estes são revistos de forma a garantir que os objetivos do negócio serão cumpridos (Chapman et al, 2000).

Por último, tem-se a fase da implantação. A criação do modelo geralmente não é o fim do negócio. É nesta fase que o modelo que foi desenvolvido pelo projeto é integrado nos

processos da organização. Assim sendo, todo o conhecimento adquirido será organizado com o intuito de se tornar viável a utilização por parte do cliente (Chapman et al, 2000).

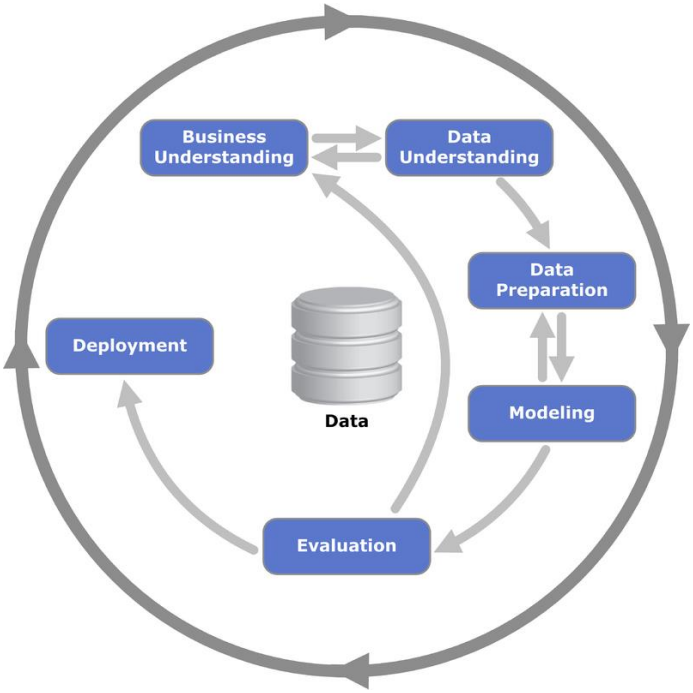


Figura 2 - Fases do modelo CRISP-DM

Fonte: Chapman et al, 2000, pp. 10

2.2.2. Modelos do *Data Mining*

Após uma visão geral do *Data Mining* e das suas vantagens, será pertinente identificar os diversos tipos de *Data Mining* existentes.

A análise exploratória de dados é uma abordagem que utiliza uma variedade de técnicas para obter uma noção inicial dos dados existentes. Dessa forma, utiliza técnicas de carácter mais visual e descritivo. Um bom exemplo desta técnica é a realização de um gráfico de dispersão de dados definidos no plano e a posterior análise visual das características desse conjunto de dados. Se for possível aproximar esse conjunto de pontos usando uma reta, podemos caracterizar este conjunto de dados como lineares. (Vaisman & Zimányi, 2014).

Por sua vez, o modelo descritivo tem como objetivo descrever os dados. Para tal, utiliza diversas técnicas. Estas, comparam os dados e agrupam-nos com o intuito de tornar a análise mais simples e percetível. O número de grupos que o modelo irá criar pode ser definido pelo modelo ou através de um algoritmo (Vaisman & Zimányi, 2014).

Seguidamente, o modelo de previsão visa a construção de um modelo que prevê o valor de uma variável a partir de valores de outras variáveis. As técnicas incluídas neste modelo irão, através dos dados existentes, criar ligações entre eles e, assim, definir uma nova variável correspondente aos dados utilizados. A principal diferença entre a previsão e a descrição é que a previsão tem uma variável como objetivo, enquanto em problemas descritivos nenhuma variável individual é fundamental para o modelo (Vaisman & Zimányi, 2014).

Por fim, a descoberta de padrões visa demonstrar comportamentos regulares e sequenciais criando um padrão de comportamento. Esta técnica é de extrema relevância não quando esse padrão se mantem inalterado, mas sim quando ocorre o oposto. Ou seja, quando existe um desvio ao comportamento normal. Através deste modelo, é mais simples perceber desvios avultados em movimentos de dinheiro que, não raras vezes, estão relacionados com corrupção e atividade fraudulenta (Vaisman & Zimányi, 2014).

Concluindo, cada um destes modelos utiliza técnicas distintas. Porém, todas elas têm em comum a finalidade de organizar os dados recolhidos. Apesar da finalidade comum, cada uma das técnicas irá dispor a informação de forma diferente.

2.2.3. Principais técnicas de *Data Mining*

Segundo Berry e Linoff (2004), o método mais utilizado e o mais comum é o de classificação. Esta técnica consiste em ordenar dados agrupando-os em classes previamente definidas. Essas classes irão tomar valores qualitativos. Normalmente, os dados estão organizados numa tabela na qual se acrescenta uma coluna com a categoria a que cada dado pertence. Por fim, esta técnica é bastante utilizada nos modelos de previsão.

Seguidamente, outra técnica frequentemente utilizada nos modelos de previsão, é a regressão. No fundo, esta técnica assemelha-se bastante à classificação, no entanto, apresenta uma diferença clara. Enquanto a classificação faz uma avaliação qualitativa, tendo em conta que classifica os dados em grupos, normalmente de sim ou não, neste caso, a avaliação é quantitativa, ou seja, cada objeto de estudo irá ter um valor provável, sendo assim possível perceber tendências individuais de forma muito mais precisa. Por outras palavras, fazem-se avaliações de variáveis contínuas, tais como, o tempo, ou a altura, por exemplo. Uma vantagem desta técnica é permitir que construa uma ordem crescente ou decrescente através dos valores verificados (Berry & Linoff, 2004).

Por fim, outra das mais importantes técnicas de *Data Mining* existentes é o C*lustering*. Neste caso, mais do que nos modelos de previsão, utiliza-se nos modelos de descrição já referidos. O *Clustering* consiste no agrupamento de dados que, eventualmente, não tenham nenhuma relação entre si. Esta técnica será abordada com mais ênfase no próximo capítulo (Berry & Linoff, 2004).

2.3. *Clustering*

Após uma revisão geral do *Data Mining* e das diversas categorias que engloba, é pertinente analisar de um modo mais detalhado o C*lustering*. Com efeito, esta será a técnica escrutinada ao longo deste trabalho e sobre a qual este estudo se irá debruçar.

Tal como já foi referido, existem diversas técnicas de organização e de abordagem de dados, todas elas com o intuito de facilitarem a compreensão do utilizador. O método de C*lustering*, como o próprio nome indica, consiste na divisão de dados em grupos que, entre si, podem não ter nenhum tipo de hierarquia nem semelhança (Witten & Frank, 2000).

Tal como na classificação, também neste caso se dividem os dados em classes. Porém, enquanto na primeira existe uma relação entre as classes, no caso do C*lustering* essa relação é bastante diminuta ou até mesmo inexistente. Na realidade, o agrupamento de dados existente nos clusters agrupa dados com grandes semelhanças entre si, mas com grandes diferenças entre grupos (Witten & Frank, 2000).

Normalmente, este tipo de método é bastante útil em análises com bases de dados muito grandes que, à partida, terão maior diversidade de conteúdos e dados mais dispares entre si. Noutro sentido, o *Clustering* não costuma ser usado como método isolado. Não é um fim em si mesmo mas sim um meio, normalmente uma iniciação, para uma abordagem mais completa dos dados (Berry & Linoff, 1997).

Apesar de ser um método bastante eficaz, também os *clusters* podem ser avaliados e classificados entre si, existindo formas de perceber o que é um cluster completo e que demonstra uma visão real dos dados existentes e aqueles nos quais essa organização é mais incompleta (Vaisman & Zimányi, 2014).

Desde logo, visualizando um gráfico de um *cluster*, consegue-se constatar que, quanto mais aglomerados forem os dados dentro do cluster, mais interessante é a análise. Por outras palavras, uma grande aproximação entre os dados de um *cluster* significa que existem grandes semelhanças entre eles, sendo esse mesmo o objetivo da divisão dos dados (Vaisman & Zimányi, 2014).

Em sentido oposto, pretende-se que a distância entre *clusters* seja grande e bem vincada. Esta situação demonstra a existência de diferenças claras e facilmente visíveis entre os dados de diferentes grupos, o que, por si, justifica que pertençam a grupos diferentes (Vaisman & Zimányi, 2014).

Para se entender melhor este conceito, basta conceber a imagem de um mapa de dados em que pontos de diferentes *clusters* se cruzem. Acontecendo este facto, torna-se injustificável que esses dados pertençam a *clusters* diferentes visto que têm efetivamente semelhanças entre si, não sendo esse o objetivo deste método de divisão e organização de dados (Vaisman & Zimányi, 2014).

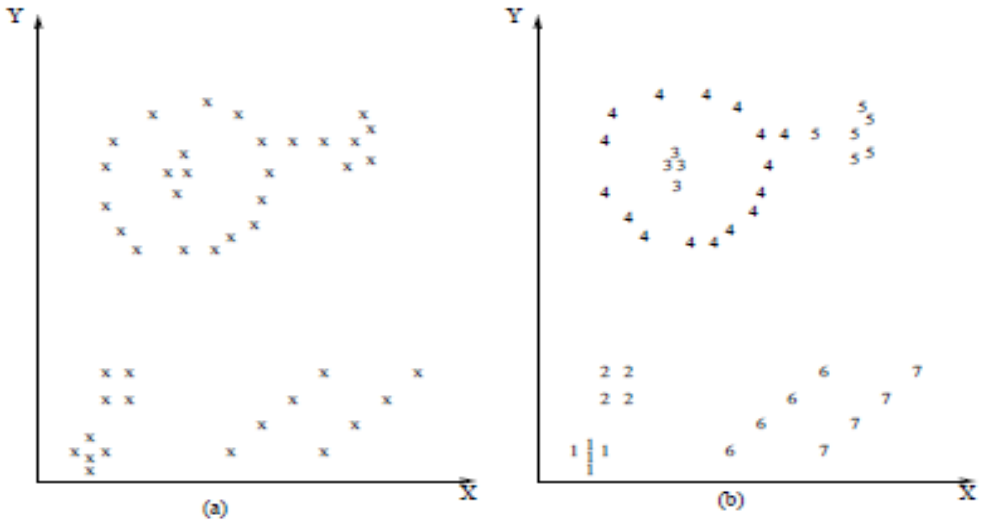


Figura 3 - *Data Clustering*
Fonte: Jain & Flynn, 1999, pp. 266

Um exemplo de agrupamento é representado na Figura 3. Sendo que os padrões de entrada são mostrados na figura 3 (a) e os agrupamentos desejados são mostrados na figura 3 (b). Assim, os pontos pertencentes ao mesmo grupo apresentam o mesmo número (Jain & Flynn, 1999)

2.3.1. Diferentes abordagens de *Clustering*

Tal como o *Data Mining* em geral, também o C*lustering* pode ser abordado tendo por base diversas técnicas. Na verdade, dependendo da finalidade dada aos dados analisados, irá variar o tipo de abordagem aplicada dentro do C*lustering*. Consequentemente, os algoritmos construídos para analisar os dados existentes, irão variar consoante o objetivo e as características do estudo que se está a analisar. Inicialmente, é primordial dividir o C*lustering* em dois grupos: os C*lustering* hierárquico e o não hierárquico (Jain, 2010).

Nos primeiros, tal como o nome indica, a forma de abordagem dos dados é hierárquica. Toma-se como ponto de partida um ponto de *cluster* e, desse, encontram-se os mais próximos. Assim, de modo crescente, vão-se formando novos *clusters*. Outro modo, também ele hierárquico, consiste em analisar todos os pontos de um grande *cluster* em simultâneo e, a partir deste, criar clusters mais pequenos (Jain, 2010).

Em relação ao C*lustering* não hierárquico, consiste em analisar os dados de uma forma geral e sem nenhuma sequência específica. Neste método, não se comparam os pontos apenas com os mais próximos mas sim todas as combinações possíveis. Através de uma avaliação geral, agrupam-se os dados de modo a criar *clusters* sem nenhuma lógica hierárquica. Naturalmente, este é um método mais complexo e bastante mais demorado (Jain, 2010).

Analisando mais detalhadamente os métodos não-hierárquicos, neste caso, os dados são decompostos em k *clusters*, sendo que cada um deverá satisfazer a maioria dos padrões de avaliação. Para obter melhores soluções e agrupamentos mais eficazes, temos de observar todas as ligações possíveis entre os dados. Um dos métodos não hierárquicos mais utilizados é o método do *k-means*. (Motoyoshi & Shioya, 2003).

2.4. Indicadores de Performance

Em circunstâncias normais, os dados são utilizados com diversos intuitos e ajudam os utilizadores a retirar novas conclusões e a melhorar o seu desempenho ou, eventualmente, de toda uma empresa. Porém, muitas vezes, é apesentada uma visão bastante redutora do fenómeno estudado (Vaisman & Zimányi, 2014).

Através das KPI's, (*Key Performance Indicators*) torna-se possível avaliar se os indicadores analisados representam de uma forma competente o problema estudado. Por outras palavras, as KPI's avaliam as variáveis utilizadas, permitindo avaliar um problema com recurso a mais variáveis que influenciam os resultados obtidos (Vaisman & Zimányi, 2014).

Normalmente, os gestores usam ferramentas de *reporting* de forma a exibir modelos estatísticos com objetivo de mostrar o desempenho de uma organização. A título de exemplo, estes relatórios podem exibir o número de alunos que se formaram na Universidade do Porto, os alunos que se formaram pela Faculdade de Economia e Gestão ou os alunos que se formaram em Economia (Vaisman & Zimányi, 2014).

Uma das operações associadas aos indicadores de performance são as operações de *drill-down* e *roll-up* que estão representadas no exemplo dado. O *drill-down* consiste na associação a uma informação geral de informação cada vez mais detalhada, consoante o interesse do utilizador saber sobre o assunto estudado. Por sua vez, o *roll-up* trata-se do desdobramento inverso. Ou seja, partindo de informação bastante pormenorizada, para uma informação mais geral acerca do mesmo caso de estudo (Vaisman & Zimányi, 2014).

As KPI's irão permitir avaliar a performance dos indicadores utilizados e, assim, dar uma visão mais global do problema em si, no caso referido, o sucesso escolar dos alunos do ensino superior da Universidade do Porto. Resumindo, os KPI's são medidas utilizadas para estimar a eficácia de uma organização no exercício das suas atividades. Desta forma, é possível monitorizar o desempenho dos processos e estratégias do negócio (Vaisman & Zimányi, 2014).

Também os KPI's têm diversas formas de serem classificados podendo sê-lo simplesmente tendo em conta o setor de atividade em que irão ser utilizados (empresas

de engenharia, empresas da área da saúde, por exemplo). Noutro âmbito, podem ser classificados de acordo com a área (KPI's para recursos humanos, KPI's para o setor financeiro). Estes são só alguns dos exemplos das formas possíveis de classificar os KPI's (Vaisman & Zimányi, 2014).

2.5. Regressão Linear Múltipla

Uma regressão linear consiste num modelo no qual existe uma variável dependente que, por sua vez, será explicada por diversas variáveis explicativas ou independentes. Este tipo de modelos é comumente utilizado em previsão. Um dos principais objetivos que se pretende atingir com as regressões lineares é perceber que variáveis independentes influenciam a variável dependente e em que medida o fazem. Caso o modelo apenas apresente uma variável independente, será uma regressão linear simples, por seu turno, um modelo com duas ou mais variáveis independentes será uma regressão linear múltipla (Armstrong & Brodie, 1999).

Nos exemplos posteriores, é possível distinguir-se uma regressão linear simples e uma múltipla.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (1)$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \varepsilon_i \quad (2)$$

Nos casos apresentados, y_i designa a observação da variável dependente para o i-ésimo indivíduo. Por sua vez, x_i representa o mesmo que y_i mas para as diversas variáveis independentes que influenciam o modelo. Cada uma das variáveis x_i, tem associado um parâmetro β que se trata do coeficiente de regressão. Este coeficiente permite definir a dimensão da relação entre a variável dependente associada e a variável independente. Por fim, o modelo é constituído por uma componente, ε_i, que representa o erro aleatório. É assumido que este erro é independente e que tem distribuição normal com média zero e variância desconhecida.

É igualmente importante salientar que o modelo é denominado por linear, devido à sua linearidade em relação aos parâmetros β.

Esta abordagem e utilização das regressões lineares é útil e auxilia na análise de determinados fenómenos e variáveis. (Armstrong & Brodie, 1999).

2.6. Coeficiente de determinação

Como já foi explicado anteriormente, um modelo de regressão linear demonstra uma relação entre diversas variáveis independentes (ou explicativas) e uma variável dependente. Porém, após a elaboração do modelo, torna-se pertinente avaliar a qualidade do ajustamento do mesmo. Assim sendo, é necessário perceber se as variáveis utilizadas explicam, efetivamente, a variável independente e em que medida o modelo é fiável.

Um dos métodos utilizados para avaliar a qualidade do modelo de regressão, trata-se do cálculo do Coeficiente de Determinação, R^2**.** Este coeficiente irá variar entre 0 e 1, sendo que, quanto mais próximo de 1, maior será a qualidade do modelo. Quer isto dizer que, caso $R^2 = 1$, as variações da variável independente serão inteiramente explicadas pelo modelo em causa.

O coeficiente de determinação é dado pela seguinte equação:

$$R^2 = \frac{SQR}{SQT} \tag{3}$$

Desta forma, é percetível que se trata da razão entre a soma dos quadrados explicada pela regressão e a soma dos quadrados total. Apesar da aparente simplicidade do cálculo, torna-se por demais evidente que será essencial calcular os valores para o numerador e para o denominador do modelo.

Assim sendo, para calcular a soma total dos quadrados, utiliza-se a seguinte equação:

$$SQT = \sum_i (y_i - \bar{y})^2 \tag{4}$$

Ou seja, a soma dos quadrados das diferenças entre a média das observações $\bar{y}$ e cada valor observado y_i.

Por sua vez, para calcular a soma dos quadrados da regressão, utiliza-se a fórmula seguinte:

$$SQR = \sum_{i} (f_i - \bar{y})^2 \quad (5)$$

Esta, indica a soma dos quadrados entre a diferença da média das observações com o valor estimado para cada observação.

3. ESTUDO DE CASO

3.1. Metodologia de Investigação e Dados Utilizados

Este estudo tem como intuito facilitar a análise e, posteriormente, a tomada de decisão na gestão. Tendo em conta o elevado volume de informação relativamente à atividade económica de uma empresa, com este estudo pretende-se apresentar um exemplo que mostra como este processo poderá ser automatizado.

Assim, neste capitulo iremos tentar desenvolver uma nova metodologia com a aplicação de agregações (C*lustering*) de Regressões. A justificação para esta escolha está relacionada com o facto de, na maioria dos casos, os indicadores baseados em dados quantitativos esconderem discrepâncias e não serem uma representação fidedigna da realidade. Posto isto, em vez de utilizar os indicadores referidos, o estudo irá utilizar indicadores baseados em modelos de regressão. Utilizando um exemplo, com este tipo de indicadores na área da gestão de alunos numa universidade podemos imaginar que a média de uma turma é influenciada pela média de entrada na faculdade e das atividades extracurriculares, ou seja, os resultados da turma podem ser explicados por determinadas variáveis socioeconómicas. É também importante salientar que os dados apresentados tanto podem ser trabalhados a nível individual, de curso, faculdade ou, até mesmo, a nível da Universidade.

A finalidade em usar regressões como indicadores de sistema de *Business Intelligence* em vez de números, prende-se na forma como poderemos fazer operações de agregação (*roll-up*) ou *drill-down*, por exemplo. Para tal ser possível, pretende-se explorar a possibilidade de usar *Clustering* para agregar os modelos de regressão usados como indicadores de performance. Isto sucede porque, sem a utilização de *Clustering*, não seria humanamente possível analisar os resultados obtidos com as regressões. Por outras palavras, o *Clustering* permite agrupar as regressões de modo a tornar os dados mais percetíveis para análise.

No estudo de caso aqui utilizado é utilizada uma base de dados com dados da Pordata (2016). Tal como já foi mencionado, um dos pontos primordiais para se tornar possível a aplicação do modelo estudado, é a possibilidade dos dados terem uma relação hierárquica

entre si. Desta forma, escolheu-se uma base de dados, que apresenta valores para os indicadores em vários patamares. Neste caso, os dados distribuem-se hierarquicamente, sendo o patamar mais elevado o nível nacional e o mais detalhado o de municípios, tendo outros níveis intermédios, como as NUTS II e NUTS III.

As variáveis utilizadas neste caso e cujos dados se obtiveram da base de dados já referida foram o Poder de Compra Concelhio, a Taxa de Desemprego e o Volume de Negócios das Empresas Não Financeiras no ano 2011. O primeiro indicador referido é aquele que é explicado, em parte, pelos dois outros indicadores. Mais à frente, estes três indicadores serão explicados de forma mais detalhada.

A aplicabilidade da regressão linear com o *Clustering* em *Business Intelligence* é algo nunca previamente estudado.

3.2. Gestão de Dados

Como ponto inicial, é importante distinguir análise de nível de decisão. Para a criação de modelos e automatismos que, como foi referido, tornem mais simples a interpretação dos dados, é necessário que exista um modelo de análise. Neste estudo o modelo de análise abordado e analisado é a agregação de regressões semelhantes entre si. Este método, independentemente do caso estudado, será utilizado como meio de organização de dados.

Noutro âmbito, o nível de decisão é algo que se incorpora na análise. Por outras palavras, a análise será a mesma, no entanto, poderá ser observada em diferentes níveis de decisão. Dependendo do caso estudado, o decisor poderá ter preferência na análise e comparação a um nível mais específico ou, por sua vez, a um nível mais abrangente. Aplicando esta situação a um caso concreto, ao longo deste trabalho irá estudar-se o desempenho empresarial em diferentes localizações geográficas. Assim, a análise será sempre realizada da mesma forma, no entanto, o nível de decisão tanto poderá ser a nível de municípios, o que permite uma observação mais detalhada. Ou, noutra perspetiva, a nível de NUTS I, sendo, assim, uma observação mais abrangente que apenas se agrupa em Continente, Região Autónoma dos Açores e Região Autónoma da Madeira.

3.2.1. Agregação de Indicadores de Performance Somáveis

De forma a explorar estas questões de investigação foi elaborada uma base de dados com dados retirados da Pordata (2016). Esta base de dados é composta por quatro variáveis quantitativas. Estas variáveis não são nada mais do que indicadores relevantes na avaliação da performance de empresas não financeiras. Os valores em causa estão organizados em função da localização geográfica da sede da empresa. Deste modo, é possível observar os dados tendo em conta diferentes níveis de abrangência, tais como: Portugal, NUTS I, NUTS II, NUTS III, Municípios e Ilhas. Como se pode constatar, os níveis apresentados têm uma ordem decrescente entre si, no entanto, na base de dados de onde foram retirados, essa relação não é intuitiva. Na verdade, a criação dessa relação hierárquica é um dos primeiros passos deste trabalho.

Primeiramente, será demonstrado qual o processo habitual sempre que é necessário fazer agregações de indicadores somáveis, ou seja, indicadores meramente quantitativos. Só posteriormente iremos abordar a questão cientifica deste trabalho: como podemos fazer agregações de indicadores mais complexos, neste caso, agregações de regressões.

Assim sendo, este estudo terá como objetivo tornar mais simples a análise e a tomada de decisões das instituições, percebendo através das regressões quais as variáveis económicas que terão um maior impacto em determinada região.

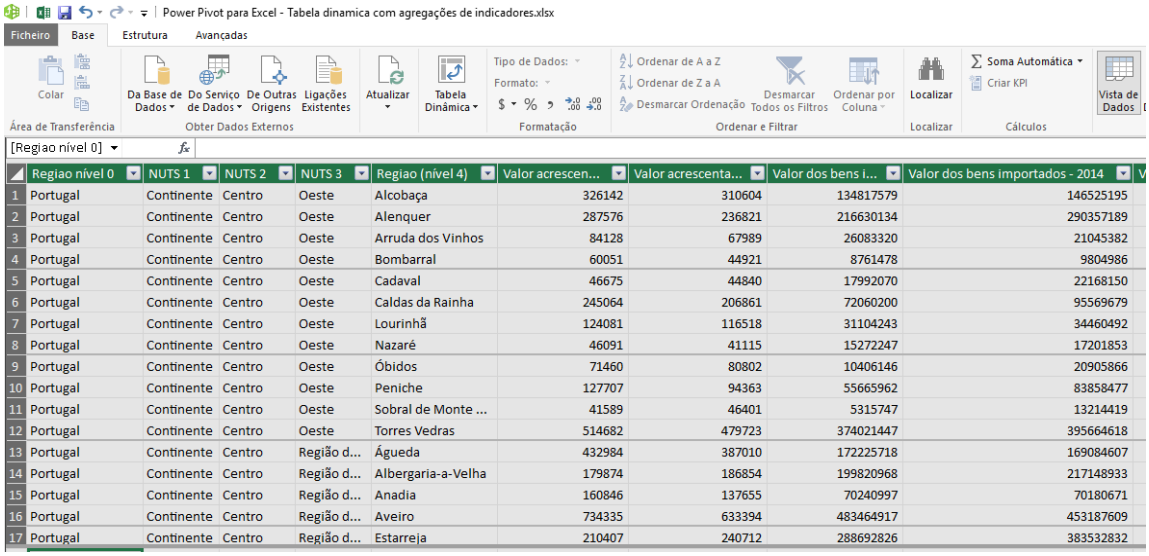
Os quatro indicadores usados na demonstração de como podemos agregar indicadores somáveis foram: o valor acrescentado bruto das empresas não financeiras; o valor dos bens importados; o valor dos bens exportados e o volume de negócios das empresas não financeiras. A seleção desses indicadores teve como principal foco perceber em que medida as diferentes variáveis que quantificam os resultados das empresas não financeiras podem impulsionar determinada região na qual as empresas se localizam. Estes dados são relativos aos anos 2010 e 2014. Deste modo, torna-se possível verificar qual a evolução temporal dos indicadores de performance das empresas em questão.

O primeiro indicador usado na base de dados criada foi o Valor Acrescentado Bruto (VAB). Este, mede a riqueza gerada na produção, descontando o valor dos bens e serviços consumidos para a obter, como são exemplo as matérias-primas.

Seguidamente, o indicador utilizado na base de dados referida foi o Valor dos Bens Importados. Este, irá contrabalançar com outro indicador, nomeadamente, o Valor dos Bens Exportados. Estes valores resultam da junção de dados provenientes de dois setores. Por um lado, do Comércio Extra-UE, que tem por base os dados recolhidos pelas alfândegas no âmbito do controlo alfandegário efetuado às transações de bens com Países Terceiros. Por outro, do Comércio Intra-UE que resulta da recolha de informação diretamente junto das empresas.

O último indicador referido é o Volume de Negócios. Este indicador caracteriza-se pelo montante obtido por uma empresa com a venda de bens e a prestação de serviços, excluindo os impostos.

Após a descrição dos dados, é importante perceber que estes indicadores são disponibilizados individualmente no site da Pordata (2016). Por conseguinte, é necessário compilar todos os indicadores recolhidos e respetivos anos, e mais importante ainda, criar a respetiva hierarquia das regiões (dimensão geográfica). Para tal, foi associado a cada região a sua região Pai. Realizou-se esta tarefa sucessivamente até se obter quatro níveis de agrupamento de dados: Portugal, NUTS I, NUTS II, NUTS III e municípios. Depois, através da ferramenta do Excel que é usada em tarefas avançadas de análise de dados e para criar modelos de dados sofisticados, designada de *Power Pivot*, os dados são importados para este modelo (Figura 4). Assim, como demonstrado na Figura 5, cria-se uma hierarquia sobre as cinco primeiras colunas.



	Regiao nível 0	NUTS 1	NUTS 2	NUTS 3	Regiao (nível 4)	Valor acrescen...	Valor acrescenta...	Valor dos bens i...	Valor dos bens importados - 2014
1	Portugal	Continente	Centro	Oeste	Alcobaça	326142	310604	134817579	146525195
2	Portugal	Continente	Centro	Oeste	Alenquer	287576	236821	216630134	290357189
3	Portugal	Continente	Centro	Oeste	Arruda dos Vinhos	84128	67989	26083320	21045382
4	Portugal	Continente	Centro	Oeste	Bombarral	60051	44921	8761478	9804986
5	Portugal	Continente	Centro	Oeste	Cadaval	46675	44840	17992070	22168150
6	Portugal	Continente	Centro	Oeste	Caldas da Rainha	245064	206861	72060200	95569679
7	Portugal	Continente	Centro	Oeste	Lourinhã	124081	116518	31104243	34460492
8	Portugal	Continente	Centro	Oeste	Nazaré	46091	41115	15272247	17201853
9	Portugal	Continente	Centro	Oeste	Óbidos	71460	80802	10406146	20905866
10	Portugal	Continente	Centro	Oeste	Peniche	127707	94363	55665962	83858477
11	Portugal	Continente	Centro	Oeste	Sobral de Monte ...	41589	46401	5315747	13214419
12	Portugal	Continente	Centro	Oeste	Torres Vedras	514682	479723	374021447	395664618
13	Portugal	Continente	Centro	Região d...	Águeda	432984	387010	172225718	169084607
14	Portugal	Continente	Centro	Região d...	Albergaria-a-Velha	179874	186854	199820968	217148933
15	Portugal	Continente	Centro	Região d...	Anadia	160846	137655	70240997	70180671
16	Portugal	Continente	Centro	Região d...	Aveiro	734335	633394	483464917	453187609
17	Portugal	Continente	Centro	Região d...	Estarreja	210407	240712	288692826	383532832

Figura 4 - Importação dos dados para a *Power Pivot*

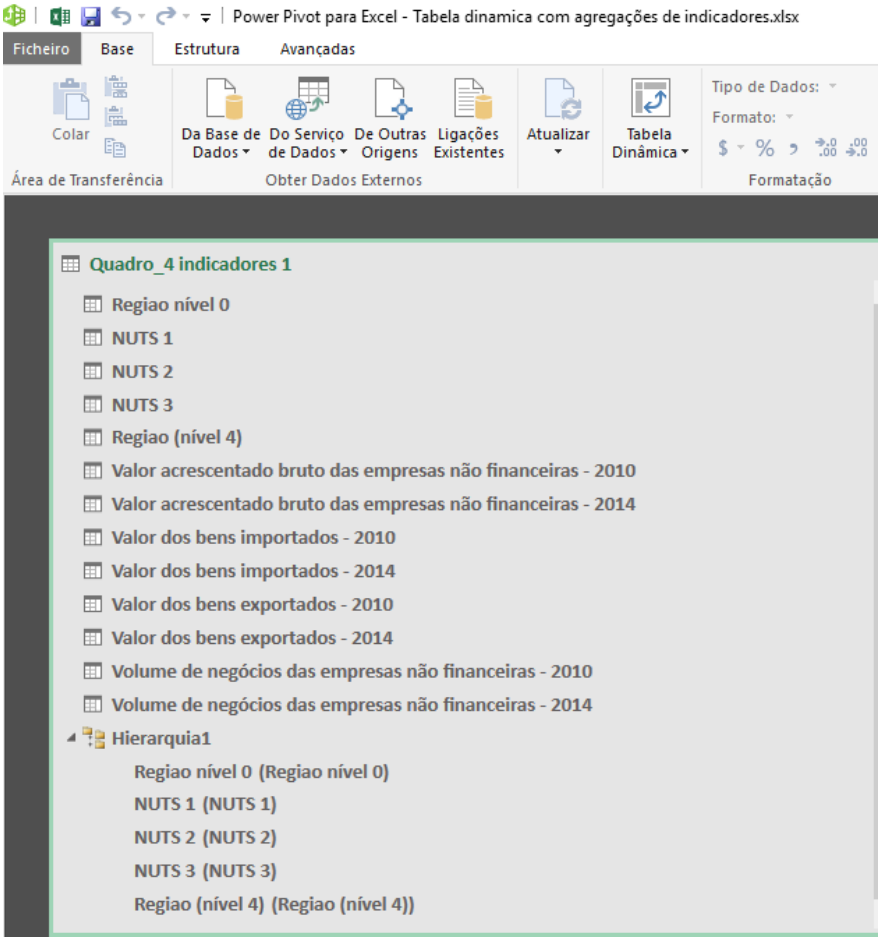


Figura 5 - Hierarquização das regiões

Por fim, irá proceder-se à criação de uma tabela dinâmica na qual as linhas da tabela terão por base a hierararquia geográfica referida anteriormente e as colunas serão compostas pelos diferentes indicadores. Esta tabela irá originar a agregação dos indicadores. Deste modo, através da utilização das operções de *drill-down* e *roll-up* torna-se possível tomar decisões com base no nível pertendido pelo utilizador. De forma mais concreta e a título de exemplo, é possível comparar o valor acrescentado bruto das empresas entre o Continente e as Regiões Autónomas. Porém, através de uma operação de *drill down*, é igualmente possível que esta informação seja desagregada. Assim, a análise poderá eventualmente fazer-se entre uma comparação dos valores do mesmo indicador, mas ao nível específico dos diferentes municípios da região Norte do país.

Logicamente, estando o modelo realizado de forma correta, o somatório de um nível de decisão, será sempre igual ao valor do nível de decisão superior. Aplicando mais uma vez

ao caso concreto, podemos verificar através da Figura 6 que o somatório do valor acrescentado bruto dos municípios que englobam a região Norte será igual ao valor da região em causa, no ano de 2010.

Regiões	Soma de Valor acrescentado bruto das empresas* (2010)	Soma de Valor acrescentado bruto das empresas (2014)	Soma de Valor dos bens importados (2010)	Soma de Valor dos bens importados (2014)	Soma de Valor dos bens exportados (2010)	Soma de Valor dos bens exportados (2014)	Soma de Volume negócios das empresas (2010)	Soma de Volume negócios das empresas (2014)
Portugal	**87466528**	**77931685**	**55730707627**	**55170810621**	**35702167755**	**46302788131**	**359606456**	**331158397**
Continente	**82124392**	**73718892**	**54954466147**	**54633717019**	**35419296949**	**45860508361**	**338562118**	**314115876**
Alentejo	3625187	3217774	2229488989	2219339052	2266227001	2877370681	14756770	14561245
Algarve	2254616	1906103	247050253	219018732	130289619	141634458	7604975	6511908
Área Metropolitana de Lisboa	40171752	34997449	33873773289	31909726170	11153492759	15364478760	164578697	150421920
Centro	13272376	11936711	6492335428	7390784927	7829773435	9251960636	55040341	52699099
Norte	22800461	21660855	12111818188	12894848138	14039514135	18225063826	96581335	89921704
Alto Minho	1094435	1183993	884055493	986505377	1061099539	1539710964	4592535	4758241
Alto Tâmega	337761	269796	42172832	42668251	31909144	59964543	1010524	909848
Área Metropolitana do Porto	13643972	12464079	8076607899	8144028550	7430021457	9348833750	60177736	53723396
Ave	2552230	2681784	1407418946	1751354812	2610109270	3503522519	9788414	10365736
Cávado	2460825	2298670	927105626	873121396	1496891262	1683313485	10515127	9363024
Douro	594038	549524	76171926	139482910	49878685	87713530	2184480	2227292
Tâmega e Sousa	1806599	1923367	470385991	614288605	1092412609	1606314258	6575127	7058712
Terras de Trás-os-Montes	310601	289642	227899475	343398237	267192169	395690777	1737392	1515455
Região Autónoma da Madeira	**2899218**	**2382831**	**315144268**	**264050516**	**116706898**	**251544268**	**10058489**	**8091837**
Região Autónoma dos Açores	**2442918**	**1829962**	**461097212**	**273043086**	**166163908**	**190735502**	**10985849**	**8950684**
Total Geral	**87466528**	**77931685**	**55730707627**	**55170810621**	**35702167755**	**46302788131**	**359606456**	**331158397**

Tabela 1 - Indicadores agrupados por localização geográfica

3.2.2. Agregação de regressões

Neste capítulo, irá ser abordado um dos pontos essenciais deste estudo. Tal como já foi referido, é importante criar indicadores de performance o mais completo possíveis. Quanto mais completo for o modelo, mais facilmente se irá detetar o que poderá estar a influenciar a variável em estudo.

Para tal, foi criada uma nova base de dados (Pordata, 2016) na qual estão contemplados indicadores que não são somáveis, como por exemplo rácios ou percentagens. Desta forma, os indicadores selecionados foram o Poder de compra concelhio, a Taxa de desemprego e o Volume de negócios das empresas não financeiras no ano 2011 (Tabela 2).

O primeiro indicador utilizado pretende traduzir o poder de compra[1] em termos per capita. Assim sendo, é um número índice com o valor 100 na média do país, que compara o poder de compra manifestado quotidianamente, em termos per capita, nos diferentes municípios ou regiões. É essencial ter em conta que na comparação temporal deste indicador, existem valores distintos que podem derivar de uma efetiva variação do poder de compra em relação à média nacional ou, por sua vez, podem também resultar de outros fatores, tais como, a utilização de um conjunto de variáveis de base na construção do indicador não totalmente coincidentes.

Seguidamente, o indicador da taxa de desemprego representa o número de desempregados por cada 100 ativos. Os ativos são a mão-de-obra disponível para trabalhar, incluindo-se na população ativa os trabalhadores que estão empregados e desempregados. Assim, esta taxa permite definir o peso da população desempregada sobre o total da população ativa.

Por último, o volume de negócios das empresas não financeiras representa a quantia líquida das vendas e prestações de serviços respeitantes às atividades normais das entidades, consequentemente após as reduções em vendas e não incluindo nem o imposto

[1] Formula matemática utilizada no cálculo do poder de compra:
IPC = (1 + CV*Factor1) / (1 + CV*FACT1Pond) *100

CV = Coeficiente de variação escolhido; Factor1 = Valores do 1º fator extraído do modelo; FACT1Pond = Valor resultante da soma para todos os municípios [Soma(Factor1)*(peso populacional)]

sobre o valor acrescentado nem outros impostos diretamente relacionados com as vendas e prestações de serviços.

Territórios		Poder de compra (%)	Taxa de desemprego segundo os Censos (%)	Volume de negócios das empresas não financeiras
Âmbito Geográfico	Anos	2011	⊥ 2011	2011
NUTS 2013	Portugal	100	13,2	341046330
NUTS I	Continente	100,8	13,2	331079748
NUTS II	Norte	89,2	14,5	93726328
NUTS III	Alto Minho	77,6	11,8	4569223
Município	Arcos de Valdevez	62	10,2	208041
Município	Caminha	81,8	13,1	167827
Município	Melgaço	62,5	9,7	56551
Município	Monção	69,4	9,8	194746
Município	Paredes de Coura	59,3	11,1	65732
Município	Ponte da Barca	60,2	13,1	91651
Município	Ponte de Lima	65	11,9	738245
Município	Valença	85,1	12,4	354329
Município	Viana do Castelo	93,1	12,5	2298259
Município	Vila Nova de Cerveira	79,9	9,7	393841
NUTS III	Cávado	85,9	12,8	10132595
Município	Amares	68,1	14,1	303040
Município	Barcelos	72,6	12,1	2706002
Município	Braga	104,2	13,2	5584355
Município	Esposende	81,5	11,3	820573
Município	Terras de Bouro	57	17,1	43660
Município	Vila Verde	64,3	12,9	674965

Tabela 2 - Parte da base de dados utilizada na agregação das regressões

Fonte: Pordata (2016)

Desta forma, com o objetivo de explicar o poder de compra em função da taxa de desemprego e do volume de negócios das empresas não financeiras procedeu-se ao

cálculo de regressões lineares múltiplas. Para isso, usou-se o Excel como ferramenta de cálculo (Figura 7).

Para cada região foi calculada uma regressão na qual a variável dependente, ou seja, a variável que será explicada pelo modelo será o poder de compra. Este será influenciado pelas variáveis independentes que, neste caso, são a taxa de desemprego e o volume de negócios.

SUMÁRIO DOS RESULTADOS

Estatística de regressão	
R múltiplo	0,939074885
Quadrado de R	0,881861639
Quadrado de R ajustado	0,848107822
Erro-padrão	7,196919802
Observações	10

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significância*
Regressão	2	2706,4554	1353,2277	26,126279	0,00056672
Residual	7	362,56958	51,795655		
Total	9	3069,025			

	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*	*nferior 95,0%*	*uperior 95,0%*
Interceptar	59,48557123	14,368281	4,1400618	0,0043488	25,50998555	93,46115692	25,509986	93,461157
Variável X 1	-0,134976267	1,0179377	-0,132598	0,8982426	-2,542016438	2,272063903	-2,542016	2,2720639
Variável X 2	1,74037E-05	2,412E-06	7,2154955	0,0001751	1,17002E-05	2,31071E-05	1,17E-05	2,311E-05

Figura 6- Cálculo da regressão linear múltipla correspondente à Ilha da Madeira

Posteriormente, as regressões calculadas foram compiladas em três tabelas, apresentadas aqui como Tabelas 2, 3 e 4. Na tabela 2, observam-se as regressões para o nível de decisão 3 que corresponde às NUTS III. Neste nível, a base utilizada para o cálculo das regressões foi o nível imediatamente inferior, ou seja, os municípios. É de salientar que, em rigor, apenas 25 regiões fazem parte das NUTS III. Porém, algumas bases de dados, como a Pordata (2016), incluem também as Ilhas. Deste modo, será esta a nomenclatura adotada neste trabalho (Tabela 2).

Seguidamente, a segunda tabela que se apresenta (Tabela 3) é meramente um processo de *roll up* em relação à tabela já referida, visto passar-se de um nível de detalhe mais aprofundado para um mais geral. Nesta, estão apresentados os dados referentes ao nível de decisão 2, correspondente às NUTS II. Trata-se de um processo de *roll up* porque estes dados foram obtidos através dos valores dos indicadores do nível de decisão 3.

Por fim, a última tabela demonstrada (Tabela 4) é a que apresenta os indicadores de uma forma mais abrangente. Assim, trata-se do nível de decisão 1 que, logicamente, corresponde às NUTS I. Tal como na tabela anterior em relação à primeira, também neste caso, os dados são obtidos através das variáveis do nível anterior.

Através da Tabela 3 podemos verificar que, por exemplo, para a região Norte estima-se que, em média, quando a taxa de desemprego varia uma unidade o poder de compra varia negativamente 2,73. No que diz respeito à variação do volume de negócios, visto tratar-se de valores bastante mais elevados, a alteração provocada no poder de compra pela variação de uma unidade no volume de negócios das empresas não financeiras, é bastante reduzida.

Por fim, é também importante referir que determinadas regiões apenas possuem uma região no seu nível inferior. Por isso não é possível calcular a regressão correspondente a esta região com apenas um ponto. Desta forma, teremos 29 regressões para as NUTS III, 4 para as NUTS II e apenas uma regressão para as NUTS I.

Tabela NUTS III	Alfa	Beta1	Beta2	p-value ANOVA	R^2
Alto Minho	51,77613738	1,329464432	1,0865E-05	0,122804781	0,450742757
Cávado	89,66461854	-1,910629117	6,30639E-06	0,075119086	0,821967128
Ave	85,04607334	-1,457783947	5,2061E-06	0,016782089	0,805043193
Área Metropolitana do Porto	80,22130366	-0,058305258	4,91682E-06	0,001592461	0,601621805
Alto Tâmega	57,17563847	-0,715179922	7,06327E-05	0,012154546	0,947136104
Tâmega e Sousa	52,9852663	-0,047057111	1,68741E-05	0,000408905	0,857798039
Douro	47,4071627	0,558921978	7,83392E-05	7,29729E-08	0,871797876
Terras de Trás-os-Montes	113,6094003	-4,246772306	3,84551E-07	0,271532725	0,310975114
Oeste	76,37164404	0,310101115	8,18458E-06	0,121782756	0,373679186
Região de Aveiro	84,07907286	-1,909930503	2,05334E-05	0,000558564	0,846266579
Região de Coimbra	92,36903229	-2,871321159	2,05022E-05	7,81694E-06	0,770052011
Região de Leiria	96,80336914	-2,364229938	7,12167E-06	0,004537715	0,785945231
Viseu Dão Lafões	57,48456072	0,203569902	1,94593E-05	2,04955E-05	0,859533009
Beira Baixa	49,3575767	0,743823408	4,78071E-05	0,025997031	0,912242851
Médio Tejo	42,23003471	2,994861831	1,04127E-05	0,193988156	0,279630954
Beiras e Serra da Estrela	56,37165461	0,321935764	4,07735E-05	2,73737E-05	0,826397645
Área Metropolitana de Lisboa	150,6190766	-3,688511589	1,40714E-06	1,43232E-05	0,773984217
Alentejo Litoral	128,734093	-4,854983035	3,58713E-05	0,045268057	0,954731943
Baixo Alentejo	69,5014583	-0,542296614	7,75616E-05	1,7393E-06	0,929518535
Lezíria do Tejo	66,29093462	0,155609283	3,32834E-05	0,000474066	0,852443048
Alto Alentejo	71,82659946	-0,508044985	8,0872E-05	6,20668E-06	0,864436375
Alentejo Central	84,08685521	-1,176810529	4,13023E-05	0,000683686	0,734221799
Algarve	49,91120554	1,365953645	3,3122E-05	0,000728641	0,670913695
Região Autónoma dos Açores	84,56035981	-1,702090426	1,62332E-05	0,00235051	0,530759738
Região Autónoma da Madeira	29,33938011	2,245038676	1,58079E-05	0,011534346	0,672283548
Ilha de Santa Maria	87,4	0	0	-	1
Ilha de São Miguel	61,96977753	-0,353671375	1,74902E-05	0,02662645	0,910832037
Ilha Terceira	65,23717018	0	3,2777E-05	-	1
Ilha da Graciosa	68,7	0	0	-	1
Ilha de São Jorge	51,52272205	0	0,00027776	-	1
Ilha do Pico	-38,6033878	12,97351579	0,000545814	-	1
Ilha do Faial	86,4	0	0	-	1
Ilha das Flores	56,35240334	0	0,000753184	-	1
Ilha do Corvo	63,1	0	0	-	1
Ilha da Madeira	59,48557123	-0,134976267	1,74037E-05	0,00056672	0,881861639
Ilha de Porto Santo	96,1	0	0	-	1

Tabela 3 - Conjunto de regressões para o nível 3 (NUTS III)

Tabela NUTS II	Alfa	Beta1	Beta2	p-value	R^2
Norte	106,1518215	-2,73143658	7,23443E-07	0,008394782	0,85222405
Centro	105,4897325	-2,568652984	1,29846E-06	0,014926337	0,813970725
Área Metropolitana de Lisboa *	131	0	0	-	1
Alentejo	111,46494	-2,111657228	1,11645E-06	0,070582101	0,929417899
Algarve	96,7	0	0	-	1
Região Autónoma dos Açores	49,95075923	3,274647754	-2,04471E-06	-	0,291080582
Região Autónoma da Madeira	96,80512344	0	-2,06684E-05	-	1

Tabela 4 - Conjunto de regressões para o nível 2 (NUTS II)

Tabela NUTS I	Alfa	Beta1	Beta2	p-value	R^2
Continente	57,53283608	1,893385732	2,35799E-07	0,381260914	0,618739086
Região Autónoma dos Açores	82,4	0	0	-	1
Região Autónoma da Madeira	85,1	0	0	-	1

Tabela 5 - Conjunto de regressões para o nível 1 (NUTS I)

3.2.3. *Clustering* de coeficientes de regressão

A agregação das regressões correspondentes à tabela das NUTS I por município foi feita através do cálculo de apenas dois clusters (k=2) uma vez que temos somente três NUTS I.

A Região Autónoma da Madeira pertence ao 1º cluster. Por sua vez, o Continente e a Região Autónoma dos Açores pertencem ao 2º cluster. Neste segundo caso, o impacto que a taxa de desemprego e o volume de negócios das empresas não financeiras têm no poder de compra nestas regiões é semelhante.

NUTS I por município	**Alfa**	**beta1**	**beta2**	**fit.cluster**
Região Autónoma da Madeira	29.33938	2.24503868	1.580791e-05	1
Continente	74.14291	0.09174324	2.295579e-06	2
Região Autónoma dos Açores	84.56036	-1.70209043	1.623318e-05	2

k=2

Group.	alfa	beta1	beta2
1	29.33938	2.2450387	1.580791e-05
2	79.35163	-0.8051736	9.264382e-06

Tabela 6 – Clusters de regressões das NUTS I por município

No caso da agregação das regressões das NUTS II por NUTS III foram calculados dois cenários. Um com dois clusters (k=2) e outro com três clusters (k=3). Analisando os centroides dos clusters do cenário em que k=3, podemos verificar o cluster 1 é muito próximo do cluster 2. Portanto, não faz sentido fazer *Clustering* com um número de k grande, deste modo, é mais plausível usar k=2.

NUTS II por NUTS III	alfa	beta1	beta2	fit.cluster
Região Autónoma dos Açores	49.95076	3.274648	-2.044713e-06	1
Região Autónoma da Madeira	96.80512	0.000000	-2.066841e-05	2
Norte	106.15182	-2.731437	7.234435e-07	3
Centro	105.48973	-2.568653	1.298459e-06	3
Alentejo	111.46494	-2.111657	1.116452e-06	3

NUTS II por NUTS III	alfa	beta1	beta2	fit.cluster
Região Autónoma dos Açores	49.95076	3.274648	-2.044713e-06	1
Norte	106.15182	-2.731437	7.234435e-07	2
Centro	105.48973	-2.568653	1.298459e-06	2
Alentejo	111.46494	-2.111657	1.116452e-06	2
Região Autónoma da Madeira	96.80512	0.000000	-2.066841e-05	2

k=3				**k=2**			
Group.1	alfa	beta1	beta2	Group.1	alfa	beta1	beta2
1	1 107.70216	-2.470582	1.046118e-06	1	1 104.97790	-1.852937	-4.382513e-06
2	2 96.80512	0.000000	-2.066841e-05	2	2 49.95076	3.274648	-2.044713e-06
3	3 49.95076	3.274648	-2.044713e-06				

Tabela 7 - Clusters de regressões das NUTS II por NUTS III

Na análise aos centroides dos clusters de coeficientes de regressões das NUTS II por município verifica-se que estes são muito dispersos. Por conseguinte, conclui-se que se justifica calcular k=3.

Assim sendo, teremos no grupo 1 a Área Metropolitana de Lisboa; no grupo 2 a Região Autónoma da Madeira e o Algarve, por fim, no grupo 3 o Norte, Centro, Alentejo e a Região Autónoma dos Açores.

NUTS II por município	**Alfa**	**beta1**	**beta2**	**fit.cluster**
Área Metropolitana de Lisboa	150.61908	-3.6885116	1.407137e-06	1
Região Autónoma da Madeira	29.33938	2.2450387	1.580791e-05	2
Algarve	38.18974	2.8923996	5.491866e-06	2
Norte	70.05224	-0.4360992	7.037962e-06	3
Centro	63.32714	0.3278163	1.526108e-05	3
Alentejo	77.70328	-0.7213196	4.086903e-05	3
Região Autónoma dos Açores	84.63500	-1.5229957	7.412089e-06	3

NUTS II por município	Alfa	beta1	beta2	fit.cluster
Área Metropolitana de Lisboa	150.61908	-3.6885116	1.407137e-06	1
Norte	70.05224	-0.4360992	7.037962e-06	2
Centro	63.32714	0.3278163	1.526108e-05	2
Alentejo	77.70328	-0.7213196	4.086903e-05	2
Algarve	38.18974	2.8923996	5.491866e-06	2
Região Autónoma dos Açores	84.63500	-1.5229957	7.412089e-06	2
Região Autónoma da Madeira	29.33938	2.2450387	1.580791e-05	2

k=3

	Group.1	alfa	beta1	beta2
1	1	73.92942	-0.5881495	1.764504e-05
2	2	33.76456	2.5687191	1.064989e-05
3	3	150.61908	-3.6885116	1.407137e-06

k=2

	Group.1	alfa	beta1	beta2
1	1	60.54113	0.464140	1.531332e-05
2	2	150.61908	-3.688512	1.407137e-06

Tabela 8 - Clusters de regressões das NUTS II por município

Relativamente ao cálculo de clusters das regressões das NUTS III por NUTS II, também faz mais sentido usar k=3. Isto sucede-se porque os centroides dos clusters apresentam uma significativa distância entre si. Podemos também verificar, no Anexo A, que a Ilha do Pico não tem qualquer tipo de semelhança com nenhuma das outras NUTS III. Deste modo, apenas esta região faz parte do cluster 1.

k=3

Group.1		alfa	beta1	beta2
1	1	45.65105	1.432413	NA
2	2	130.98752	-4.263422	NA
3	3	80.46604	-1.079672	NA

k=2

Group.1		alfa	beta1	beta2
1	1	97.18041	-2.160939	NA
2	2	50.40115	1.083761	NA

Tabela 9 - Clusters de regressões das NUTS III por NUTS II

3.3. Qualidade da Aplicabilidade do Estudo

De forma a medir a qualidade da aplicação do *Clustering* dos coeficientes de regressão é necessário recorrer, novamente, ao coeficiente de determinação, também denominado de R^2. Mais precisamente, será preciso calcular a variação que este irá sofrer devido à agregação. Com efeito, irá comparar-se o R^2 quando agregamos todos os coeficientes de regressão, com o R^2 individual de cada região em análise. Isto irá permitir perceber quanto se ganha ou se perde em termos de qualidade da regressão ao fazer as agregações.

Em termos de análise, pode concluir-se desde já, que ao fazer agregação a qualidade da regressão aumenta sempre. No entanto, será necessário testar a qualidade dos resultados.

Desta forma, calcula-se o R^2 para cada tabela em análise (NUTS I por município, NUTS II por NUTS III, NUTS II por municípios, NUTS III por municípios) e compara-se com o R^2 de cada região individual. Recordando que o coeficiente de determinação é dado pela razão entre a soma dos quadrados de regressão e a soma dos quadrados total.

$$R^2 = \frac{SQR}{SQT}$$

3.3.1. Qual a variação da qualidade do ajustamento quando se efetua a agregação das regressões?

3.3.1.1. R^2 da agregação das regressões relativas às NUTS I por municípios

Iniciando a análise pelas NUTS I por município, verifica-se que, neste caso, a qualidade do modelo agregado é demonstrada pelo valor de $R^2 \cong 0{,}0174$. Desta forma, entende-se que ocorre uma diminuição acentuada, se comparado com o R^2 de cada região individual (NUTS I). Posto isto, no modelo com C*lustering*, as variáveis independentes influenciam uma parte consideravelmente menor do poder de compra.

k=2

alfa	beta1	beta2	$\bar{y}$	77,3571
29.33938	2.2450387	1.580791e-05		
79.35163	-0.8051736	9.264382e-06		

fi	SQR
29,33938	2305,70555
79,35163	3,97797896
108,69101	**2309,68353**

SQT	**132886,114**
SQR	**2309,68353**
R^2	**0,01738092**

Tabela 10 - R^2 do modelo agregado (NUTS I por municípios)

Tabela NUTS I por municipio	Alfa	Beta1	Beta2	p-value = F signifiância	R^2
Continente	74,14291	0,09174324	2,29558E-06	3,27109E-27	0,35823
Região Autónoma dos Açores	84,56036	-1,70209043	1,62332E-05	0,00235051	0,53076
Região Autónoma da Madeira	29,33938	2,24503868	1,58079E-05	0,011534346	0,672284

Tabela 11 - R^2 de cada região individual pertencente às NUTS I

3.3.1.2. R^2 da agregação das regressões relativas às NUTS II por NUTS III

Passando para a análise da agregação do modelo das NUTS II por NUTS III, pode verificar-se que, ao contrário do caso anterior, existe um aumento do valor do R^2, sendo, neste caso, aproximadamente 0,9374. Posto isto, após a agregação, aumenta a percentagem do modelo que poderá ser explicada em função das variáveis explicativas.

k=2

alfa	beta1	beta2	$\bar{y}$	83,9242
60.54113	0.464140	1.531332e-05		
150.61908	-3.688512	1.407137e-06		

fi	SQR
60,54113	546,769947
150,61908	4448,20136
211,16021	**4994,97131**

SQT	**5328,28061**
SQR	**4994,97131**
R^2	**0,93744524**

Tabela 12 - R^2 do modelo agregado (NUTS II por NUTS III)

Tabela NUTS II por NUTS III	Alfa	Beta1	Beta2	p-value	R^2
Norte	106,1518	-2,73143658	7,23443E-07	0,008394782	0,852224
Centro	105,4897	-2,56865298	1,29846E-06	0,014926337	0,813971
Alentejo	111,4649	-2,11165723	1,11645E-06	0,070582101	0,929418
Região Autónoma dos Açores	49,95076	3,27464775	-2,04471E-06	-	0,291081
Região Autónoma da Madeira	96,80512	0	-2,06684E-05	-	1

Tabela 13 - R^2 de cada região individual pertencente às NUTS II

3.3.1.3. R^2 da agregação das regressões relativas às NUTS II por municípios

No caso das NUTS II por municípios, o $R^2 \cong 0{,}0548$ demonstra que a qualidade do modelo após a agregação diminui. Assim, pode-se constatar que o modelo de NUTS II por município tem um valor de R^2 que garante uma maior qualidade ao modelo antes do C*lustering.*

k=3

alfa	beta1	beta2
73.92942	-0.5881495	1.764504e-05
33.76456	2.5687191	1.064989e-05
150.61908	-3.6885116	1.407137e-06

$\bar{y}$ 77,357143

alfa(fi)	SSR
73,92942	11,74928399
33,76456	1900,31328
150,6108	5366,098285
258,30478	**7278,160849**

SQT	**132886,1143**
SQR	**7278,160849**
R^2	**0,054769912**

Tabela 14 - R^2 do modelo agregado (NUTS II por municípios)

Tabela NUTS II por municipios	Alfa	Beta1	Beta2	p-value = F signifiância	R^2
Norte	70,05224	-0,43609917	7,03796E-06	7,58696E-18	0,613213
Centro	63,32714	0,32781634	1,52611E-05	1,00059E-17	0,553841
Área Metropolitana de Lisboa	150,6191	-3,68851159	1,40714E-06	1,43232E-05	0,773984
Alentejo	77,70328	-0,72131957	4,0869E-05	6,84874E-19	0,781487
Algarve	38,18974	2,89239957	5,49187E-06	0,139612362	0,245175
Região Autónoma dos Açores	84,635	-1,52299574	7,41209E-06	0,024094947	0,354883
Região Autónoma da Madeira	29,33938	2,24503868	1,58079E-05	0,011534346	0,672284

Tabela 15 - R^2 de cada região individual pertencente às NUTS II

3.3.1.4. R^2 da agregação das regressões relativas às NUTS III por municípios

Por último, quando agregamos as NUTS III por municípios, o R^2 tem o valor de 0,0293. Também aqui se constata uma diminuição substancial da qualidade do ajustamento do modelo. Desta forma, a percentagem do modelo passível de ser explicada pelas variáveis independentes é maior no modelo com cada região individual do que no modelo com agregação.

k=3

	Group.1	alfa	beta1	beta2
1	1	45.65105	1.432413	NA
2	2	130.98752	-4.263422	NA
3	3	80.46604	-1.079672	NA

$\bar{y}$ 77,357143

fi	SSR
45,65105	1005,276324
130,98752	2876,217352
80,46604	9,665241445
257,10461	**3891,158918**

SQT	**132886,1143**
SQR	**3891,158918**
R^2	**0,029281908**

Tabela 16 - R^2 do modelo agregado (NUTS III por municípios)

Tabela NUTS III por municipios	Alfa	Beta1	Beta2	p-value ANOVA	R^2
Alto Minho	51,7761	1,3294644	1,0865E-05	0,122804781	0,45074
Cávado	89,6646	-1,9106291	6,30639E-06	0,075119086	0,82197
Ave	85,0461	-1,4577839	5,2061E-06	0,016782089	0,80504
Área Metropolitana do Porto	80,2213	-0,0583053	4,91682E-06	0,001592461	0,60162
Alto Tâmega	57,1756	-0,7151799	7,06327E-05	0,012154546	0,94714
Tâmega e Sousa	52,9853	-0,0470571	1,68741E-05	0,000408905	0,8578
Douro	47,4072	0,558922	7,83392E-05	7,29729E-08	0,8718
Terras de Trás-os-Montes	113,609	-4,2467723	3,84551E-07	0,271532725	0,31098
Oeste	76,3716	0,3101011	8,18458E-06	0,121782756	0,37368
Região de Aveiro	84,0791	-1,9099305	2,05334E-05	0,000558564	0,84627
Região de Coimbra	92,369	-2,8713212	2,05022E-05	7,81694E-06	0,77005
Região de Leiria	96,8034	-2,3642299	7,12167E-06	0,004537715	0,78595
Viseu Dão Lafões	57,4846	0,2035699	1,94593E-05	2,04955E-05	0,85953
Beira Baixa	49,3576	0,7438234	4,78071E-05	0,025997031	0,91224
Médio Tejo	42,23	2,9948618	1,04127E-05	0,193988156	0,27963
Beiras e Serra da Estrela	56,3717	0,3219358	4,07735E-05	2,73737E-05	0,8264
Área Metropolitana de Lisboa	150,619	-3,6885116	1,40714E-06	1,43232E-05	0,77398
Alentejo Litoral	128,734	-4,854983	3,58713E-05	0,045268057	0,95473
Baixo Alentejo	69,5015	-0,5422966	7,75616E-05	1,7393E-06	0,92952
Lezíria do Tejo	66,2909	0,1556093	3,32834E-05	0,000474066	0,85244
Alto Alentejo	71,8266	-0,508045	8,0872E-05	6,20668E-06	0,86444
Alentejo Central	84,0869	-1,1768105	4,13023E-05	0,000683686	0,73422
Algarve	49,9112	1,3659536	3,3122E-05	0,000728641	0,67091
Região Autónoma dos Açores	84,5604	-1,7020904	1,62332E-05	0,00235051	0,53076
Região Autónoma da Madeira	29,3394	2,2450387	1,58079E-05	0,011534346	0,67228
Ilha de São Miguel	61,9698	-0,3536714	1,74902E-05	0,02662645	0,91083
Ilha Terceira	65,2372	0	3,2777E-05	-	1
Ilha de São Jorge	51,5227	0	0,00027776	-	1
Ilha do Pico	-38,6034	12,973516	0,000545814	-	1
Ilha das Flores	56,3524	0	0,000753184	-	1
Ilha da Madeira	59,4856	-0,1349763	1,74037E-05	0,00056672	0,88186

Tabela 17 - R^2 de cada região individual pertencente às NUTS III

Como resposta à questão colocada no título deste capítulo, após esta análise, é claro constatar-se que, na generalidade dos casos, o modelo perde qualidade com a agregação. Como já foi referido individualmente, isto significa que existirá uma menor percentagem dos modelos que poderá ser explicada pelas variáveis independentes. Esta conclusão já era expectável. Com a agregação, o modelo torna os dados mais percetíveis e, portanto, mais simples para efetuar análise. Porém, naturalmente, a diminuição dos grupos do modelo torna o modelo menos aprofundado e, por isso mesmo, com uma amplitude de detalhe menor. Para efeitos de maior entendimento, no extremo, se cada dado fosse analisado individualmente, iria ter uma qualidade de explicação de 100%. Evidentemente que, em quantos menos grupos se efetuar a agregação, menor será a percentagem do modelo explicada pelas variáveis usadas.

4. CONCLUSÕES E CONSIDERASÕES FINAIS

Este trabalho tem como objetivo final o desenvolvimento de um método com vista a facilitar e tornar mais eficaz a análise de dados em grande escala que, por sua vez, não fossem passíveis de serem somados. Esse foi o ponto de partida e o fio condutor de todo o trabalho realizado ao longo deste estudo. Na verdade, desde o primeiro momento, era esse o objetivo primordial e foi sempre com vista a alcançá-lo que a investigação se guiou. Através do método utilizado, pretendeu-se tornar a análise de dados não só mais simples e eficaz como, igualmente, mais imparcial e menos suscetível à influência de fatores externos.

Tal como em muitos outros casos, também aqui, para que a simplicidade seja fornecida ao utilizador, é necessário elevar o nível de complexidade na elaboração do sistema que fornece o serviço. Nesse sentido, para que a interpretação de dados por parte de quem utiliza o método seja simplificada, foi essencial o aprofundamento do tema e realizar um estudo completo e com maior nível de detalhe.

Desde logo, ao longo da revisão de literatura pretendeu-se, numa primeira fase, clarificar e definir conceitos. Num segundo momento, foi extremamente importante a contextualização dos mesmos em função do tema estudado. Deste segundo ponto, advém a primeira grande conclusão do trabalho realizado. O facto de se tratar de um método pouco estudado, tornou a pesquisa mais complexa, tendo em conta que não existe teoria sobre o tema exato que se aborda ao longo desta dissertação. Desse modo, foi necessário conjugar estudos efetuados na área para enriquecer o conhecimento previamente a passar para a análise empírica. Esse trabalho de concretização da teoria foi deveras esclarecedor e tornou toda a análise posterior mais aliciante e menos abstrata.

Posteriormente, com a aplicação do método, concluiu-se o que a teoria antecipava. A passagem da análise de dados somáveis para a de não somáveis, aumenta claramente a complexidade do agrupamento de dados. Ao tratarem-se de dados somáveis possibilita que a técnica de *Clustering* seja aplicada de forma mais simples e garante a hipótese de efetuar o *roll up* e *drill down* pretendidos, visto que o nível superior será simplesmente o somatório dos constituintes do grupo inferior.

Por sua vez, passemos a focar a conclusão no tema central da dissertação: o *Clustering* de dados não somáveis. Aqui, constata-se que este método é realmente vantajoso, possibilitando uma análise mais completa e pertinente dos dados que, à partida, não seriam agrupáveis. Referindo a base de dados utilizada no estudo realizado, partindo do pressuposto que o Estado iria querer aplicar determinadas políticas, após a aplicação deste método, o trabalho destes iria estar facilitado. Com efeito, as regiões do país estão agrupadas de acordo com a influência que a taxa de desemprego e o volume de negócios das empresas não financeiras (as variáveis explicativas) têm no poder de compra (a variável explicada). Estes grupos podem ser constituídos por regiões do país sem qualquer tipo de ligação geográfica. O que faz com que pertençam ao mesmo grupo é simplesmente a influência que as variáveis explicativas têm na variável explicada.

Sendo aplicável em diversas áreas, este estudo tem como grande vantagem auxiliar os gestores e os órgãos de decisão das empresas e instituições. Tendo em conta o volume elevado de informação relativa à atividade económica das empresas, o processo de análise de dados e tomada de decisões de gestão é, por vezes, bastante complexo. Através deste estudo, dá-se a possibilidade ao utilizador de analisar elevadas quantidades de dados de forma mais simples, mas ao mesmo tempo, mais detalhada, podendo levar, deste modo, à tomada de melhores decisões empresariais.

Noutro âmbito, evidentemente que, em comparação com a análise de dados somáveis, neste caso, as ações de *roll up* e *drill down* não são tão intuitivas. Este facto torna-se óbvio se se clarificar que, naturalmente, os níveis superiores (inferiores) não são uma mera soma (subtração) dos níveis inferiores (superiores). Neste caso, como foi percetível ao longo do estudo, têm de se elaborar novas tabelas que permitem interpretar os dados em função do patamar que se pretende, dependendo se o utilizador quer ter uma visão mais abrangente ou mais detalhada.

Uma das claras limitações deste estudo prende-se com o número limitado de variáveis explicativas incluídas no modelo. Para efeitos de exemplificação do método elaborado, a utilização de duas variáveis explicativas é suficientemente esclarecedora, no entanto, num caso real, existem inúmeras variáveis que irão influenciar a variável explicada pelo modelo que, no caso apresentado, é o poder de compra. É precisamente por esse fator que os coeficientes de determinação apresentam valores tão pequenos, dado que este

coeficiente, como já foi explicado, confere a percentagem do modelo explicada pelas variáveis explicativas usadas.

No mesmo sentido, os níveis de análise utilizados neste estudo são, também eles, reduzidos. Mais uma vez, para efeitos de explicação do método estudado, torna-se percetível e claro para o efeito. Porém, numa situação real o estudo teria de ser substancialmente mais aprofundado indo, por exemplo, até ao agrupamento por freguesias.

Como finalização, é relevante perceber-se que este método específico, a aplicação de *Clustering* a regressões lineares múltiplas de dados não somáveis, poderá futuramente ser implementado de forma mais completa e efetiva. Deste modo, com base neste método, poderá criar-se uma nova ferramenta de *Business Intelligence*, com o intuito de agrupar indicadores mais complexos e, consequentemente, ser possível agregar e desagregar de forma automática através do *roll up* e *drill down* supramencionados.

5. BIBLIOGRAFIA

Armstrong, J. S., & Brodie, R. (1999). Forecasting for marketing.

Berry, M. J., & Linoff, G. (1997). Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc..

Berry, M. J., & Linoff, G. S. (2004). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. ACM Sigmod record, 26(1), 65-74.

Fundação Francisco Manuel dos Santos – Pordata: Base de Dados Portugal Contemporâneo (2011), http://www.pordata.pt/Municipios/Poder+de+compra+per+capita-118, acedido em 23 Junho 2016.

Fundação Francisco Manuel dos Santos – Pordata: Base de Dados Portugal Contemporâneo (2011), http://www.pordata.pt/Municipios/Taxa+de+desemprego+segundo+os+Censos+total+e+por+sexo+(percentagem)-405, acedido em 23 Junho 2016.

Fundação Francisco Manuel dos Santos – Pordata: Base de Dados Portugal Contemporâneo (2011), https://www.pordata.pt/Municipios/Valor+dos+bens+importados+e+exportados+pelas+empresas-393, acedido em 06 Junho 2016.

Fundação Francisco Manuel dos Santos – Pordata: Base de Dados Portugal Contemporâneo (2011), http://www.pordata.pt/Municipios/Volume+de+neg%C3%B3cios+das+empresas+n%C3%A3o+financeiras+total+e+por+sector+de+actividade+econ%C3%B3mica-589, acedido em 23 Junho 2016.

Fundação Francisco Manuel dos Santos – Pordata: Base de Dados Portugal Contemporâneo (2011), https://www.pordata.pt/Site/MicroPage.aspx?DatabaseName=Municipios&MicroName=Valor+acrescentado+bruto+das+empresas+n%C3%A3o+financeiras+total+e+por+sector+de+actividade+econ%C3%B3mica&MicroURL=588&, acedido em 06 Junho 2016.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

Motoyoshi, M., Miura, T., & Shioya, I. (2003). Clustering by regression analysis. In Data Warehousing and Knowledge Discovery (pp. 202-211). Springer Berlin Heidelberg.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling*. John Wiley & Sons.

Vaisman, A., & Zimányi, E. (2014). Data Warehouse Systems: Design and Implementation. Springer.

Witten, I. H., & Frank, E. (2000). Data mining: practical machine learning tools and techniques with Java implementations. 2000.

6. ANEXOS

Anexo A – Cálculo dos clusters relativos às regressões das NUTS III por NUTS II

NUTS III por NUTS II	Alfa	beta1	beta2	fit.cluster
Ilha do Pico	-38.60339	12.97351579	0.000545814405663914	1
Cávado	89.66462	-1.91062912	6.30638632264795E-06	2
Ave	85.04607	-1.45778395	5.20609602469109E-06	2
Área Metropolitana do Porto	80.22130	-0.05830526	4.91682451717678E-06	2
Terras de Trás-os-Montes	113.60940	-4.24677231	3.84551395039713E-07	2
Região de Aveiro	84.07907	-1.90993050	0.0000205333912275806	2
Região de Coimbra	92.36903	-2.87132116	0.0000205022415646569	2
Região de Leiria	96.80337	-2.36422994	7.12166642975054E-06	2
Área Metropolitana de Lisboa	150.61908	-3.68851159	0.0000014071370207221	2
Alentejo Litoral	128.73409	-4.85498304	0.0000358712823014293	2
Alentejo Central	84.08686	-1.17681053	0.0000413022760498919	2
Região Autónoma dos Açores	84.56036	-1.70209043	0.0000162331846018748	2
Alto Minho	51.77614	1.32946443	0.0000108649597692351	3
Alto Tâmega	57.17564	-0.71517992	0.0000706326946354651	3
Tâmega e Sousa	52.98527	-0.04705711	0.0000168741103954013	3
Douro	47.40716	0.55892198	0.0000783391815382582	3
Oeste	76.37164	0.31010111	8.18458222521235E-06	3
Viseu Dão Lafões	57.48456	0.20356990	0.0000194593482899757	3

Beira Baixa	49.35758	0.74382341	0.0000478071351436103	3
Médio Tejo	42.23003	2.99486183	0.000010412685478723	3
Beiras e Serra da Estrela	56.37165	0.32193576	0.0000407734954863624	3
Baixo Alentejo	69.50146	-0.54229661	0.0000775615845431022	3
Lezíria do Tejo	66.29093	0.15560928	0.0000332834340988873	3
Alto Alentejo	71.82660	-0.50804499	0.000080871958453738	3
Algarve	49.91121	1.36595364	0.0000331219501027909	3
Região Autónoma da Madeira	29.33938	2.24503868	0.0000158079080419813	3
Ilha de São Miguel	61.96978	-0.35367138	0.0000174902410360843	3
Ilha Terceira	65.23717	0.00000000	0.0000327770053447665	3
Ilha de São Jorge	51.52272	0.00000000	0.000277760040865845	3
Ilha das Flores	56.35240	0.00000000	0.000753183913817501	3
Ilha da Madeira	59.48557	-0.13497627	0.0000174036663625069	3

Tabela 18 - Clusters relativos às regressões das NUTS III por NUTS II com k=3

NUTS III por NUTS II	alfa	beta1	beta2	fit.cluster
Ilha do Pico	-38.60339	12.97351579	0.000545814405663914	1
Alto Minho	51.77614	1.32946443	0.0000108649597692351	2
Cávado	89.66462	-1.91062912	6.30638632264795E-06	2
Ave	85.04607	-1.45778395	5.20609602469109E-06	2
Área Metropolitana do Porto	80.22130	-0.05830526	4.91682451717678E-06	2
Alto Tâmega	57.17564	-0.71517992	0.0000706326946354651	2
Tâmega e Sousa	52.98527	-0.04705711	0.0000168741103954013	2
Douro	47.40716	0.55892198	0.0000783391815382582	2
Terras de Trás-os-Montes	113.60940	-4.24677231	3.84551395039713E-07	2
Oeste	76.37164	0.31010111	8.18458222521235E-06	2
Região de Aveiro	84.07907	-1.90993050	0.0000205333912275806	2
Região de Coimbra	92.36903	-2.87132116	0.0000205022415646569	2
Região de Leiria	96.80337	-2.36422994	7.12166642975054E-06	2
Viseu Dão Lafões	57.48456	0.20356990	0.0000194593482899757	2
Beira Baixa	49.35758	0.74382341	0.0000478071351436103	2
Médio Tejo	42.23003	2.99486183	0.0000010412685478723	2
Beiras e Serra da Estrela	56.37165	0.32193576	0.0000407734954863624	2
Área Metropolitana de Lisboa	150.61908	-3.68851159	0.0000014071370207221	2
Alentejo Litoral	128.73409	-4.85498304	0.0000358712823014293	2
Baixo Alentejo	69.50146	-0.54229661	0.0000775615845431022	2

Lezíria do Tejo	66.29093	0.15560928	0.0000332834340988873	2
Alto Alentejo	71.82660	-0.50804499	0.000080871958453738	2
Alentejo Central	84.08686	-1.17681053	0.0000413022760498919	2
Algarve	49.91121	1.36595364	0.0000331219501027909	2
Região Autónoma dos Açores	84.56036	-1.70209043	0.0000162331846018748	2
Região Autónoma da Madeira	29.33938	2.24503868	0.0000158079080419813	2
Ilha de São Miguel	61.96978	-0.35367138	0.0000174902410360843	2
Ilha Terceira	65.23717	0.00000000	0.0000327770053447665	2
Ilha de São Jorge	51.52272	0.00000000	0.000277760040865845	2
Ilha das Flores	56.35240	0.00000000	0.000753183913817501	2
Ilha da Madeira	59.48557	-0.13497627	0.0000174036663625069	2

Tabela 19 - Clusters relativos às regressões das NUTS III por NUTS II com k=2